A Bayesian Nonparametric Estimation to Entropy

Luai Al-Labadi*1, Vishakh Patel1, Kasra Vakiloroayaei1, and Clement Wan§1

1Department of Mathematical and Computational Sciences, University of Toronto Mississauga, Mississauga, Ontario L5L 1C6, Canada.

Abstract

A Bayesian nonparametric estimator to entropy is proposed. The derivation of the new estimator relies on using the Dirichlet process and adapting the well-known frequentist estimators of Vasicek (1976) and Ebrahimi, Pflughoft and Soofi (1994). Several theoretical properties, such as consistency, of the proposed estimator are obtained. The quality of the proposed estimator has been investigated through several examples, in which it exhibits excellent performance.

Keywords: Dirichlet process; Kullback-Leibler divergence; Model checking.

MSC 2000 62F15, 94A17, 62F03.

*Corresponding author: luai.allabadi@utoronto.ca
1vishakh.patel@mail.utoronto.ca
1k.vakiloroayaei@mail.utoronto.ca
§cmclement.wan@mail.utoronto.ca
1 Introduction

Entropy is a central quantity in information theory that finds applications in several fields such as thermodynamics, communication theory, computer science, biology, economics and statistics (Cover and Thomas, 2006). Let $X$ be a random variable with cumulative distribution function (cdf) $F$ and probability density function (pdf) $f$. The entropy $H(F)$ of $X$ is defined by Shannon (1948) as

$$H(F) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx.$$  

(1)

In most realistic applications, the true pdf $f$ is unknown. Hence, one must estimate (1) from the data, which is not a trivial task. Various frequentist procedures for the estimation of entropy are offered in the literature. Among several estimators, due to its simplicity, the estimator of Vasicek (1976) has been the most widely used one. Vasicek (1976) noticed that (1) can be expressed as

$$H(F) = -\int_{0}^{1} \log \left( \frac{d}{dt} F^{-1}(t) \right) dt.$$  

An estimate of $H(F)$ can be constructed by replacing the distribution function $F$ by the empirical distribution function $F_n$ and using a difference operator instead of the differential operator. Then the derivative of $F^{-1}(t)$ is estimated by a function of the order statistics. Specifically, if $x = (x_1, \ldots, x_n)$ is a sample from $F$, then Vasicek's (1976) estimator is given by

$$H_{m,n}^{V} = n^{-1} \sum_{i=1}^{n} \log \left( \frac{x_{(i+m)} - x_{(i-m)}}{2m/n} \right),$$  

(2)

where $m$ is a positive integer smaller than $n/2$ and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics of $x_1, x_2, \ldots, x_n$ with $x_{(i-m)} = x_{(1)}$ if $i \leq m$, $x_{(i+m)} = x_{(n)}$ if $i \geq n - m$. Vasicek (1976) showed that $H_{m,n}^{V} \xrightarrow{p} H(F)$ as $n \to \infty, m \to \infty$ and
\( \frac{m}{n} \to 0 \), where \( \overset{p}{\to} \) denotes convergence in probability. Note that, the expression inside the log in (2) is the slope of the straight line that passes through the points \( \left( \frac{i+m}{n}, x_{i+m} \right) \) and \( \left( \frac{i-m}{n}, x_{i-m} \right) \). Here, \( F_n(x_{i+m}) = \frac{i+m}{n} \) and \( F_n(x_{i-m}) = \frac{i-m}{n} \). Ebrahimi, Pflughoeft and Soofi (1994) noticed that (2) does not give the correct formula for the slope when \( i \leq m \) or \( i \geq n - m + 1 \). They proposed the following modification to (2):

\[
H_{m,n}^{EPS} = n^{-1} \sum_{i=1}^{n} \log \left( \frac{x_{i+m} - x_{i-m}}{c_i m/n} \right),
\]

(3)

where

\[
c_i = \begin{cases} 
\frac{m+i-1}{m} & \text{if } 1 \leq i \leq m \\
2 & \text{if } m + 1 \leq i \leq n - m \\
\frac{n+m-i}{m} & \text{if } n - m + 1 \leq i \leq n 
\end{cases}
\]

(4)

They also showed that \( H_{m,n}^{EPS} \overset{p}{\to} H(F) \) as \( n \to \infty, m \to \infty \) and \( \frac{m}{n} \to 0 \).

Other nonparametric frequentist estimators of entropy include, among others, the works of van Es (1992), Correa (1995), Wieczorkowski and Grzegorzewski (1999), Alizadeh Noughabi (2010), Alizadeh Noughabi and Arghami (2010) and Al-Omari (2016). We refer the reader to the work of Beirlant, Dudewicz, Győria and van der Meulen (1997) of a comprehensive review for nonparametric entropy estimators.

On the contrary to frequentist approach, the Bayesian estimation of entropy did not receive much attention. A relevant work includes that of Mazzuchi, Soofi and Soyer (2008) who develop a Bayes estimator of \( H(F) \) based on the Dirichlet process (Ferguson, 1973).

The goal of this paper is to provide a Bayesian approach to estimation of entropy. We start by reviewing the Dirichlet process prior in Section 2. In Section 3, a Bayesian nonparametric estimator of the entropy is obtained and several
of its properties are derived. Section 4 develops a computational algorithm of
the approach, where particular choices of \( m \) and the hyperparameters of the
Dirichlet process are suggested. Section 5 presents several examples to assess
the quality of the proposed estimator. Section 6 ends with a brief summary of
the results. Proofs are found in the Appendix.

2 Dirichlet process

The Dirichlet process was introduced by Ferguson in 1973 as a prior on the
space of probability measures. It is quantified by two ingredients, namely, the
*base measure* \( G \) (a fixed probability measure) and the *concentration parameter* \( a \) (a positive real number). It can be viewed as the infinite-dimensional
generalization of the Dirichlet distribution constructed around \( G \), where its
variation is controlled by \( a \). The Dirichlet process is formally defined as fol-
lows. Consider a space \( \mathcal{X} \) with a \( \sigma \)-algebra \( \mathcal{A} \) of subsets of \( \mathcal{X} \). For a base mea-
sure \( G \) on \( (\mathcal{X}, \mathcal{A}) \) and \( a > 0 \), a random probability measure \( P = \{P(A)\}_{A \in \mathcal{A}} \)
is called a Dirichlet process on \( (\mathcal{X}, \mathcal{A}) \), denoted by \( P \sim DP(a,G) \), if for every
measurable partition \( \{A_1, \ldots, A_k\} \) of \( \mathcal{X} \) with \( k \geq 2 \), the joint distribution
of the vector \( (P(A_1), \ldots, P(A_k)) \) has the Dirichlet distribution with parame-
ters \( ((aG(A_1), \ldots, aG(A_k))) \). It is assumed that \( G(A_j) = 0 \) implies \( P(A_j) = 0 \)
with probability one. From the properties of the Dirichlet distribution, for any
\( A \in \mathcal{A}, P(A) \sim \text{beta}(aG(A), a(1 - G(A))), E(P(A)) = G(A) \) and \( \text{Var}(P(A)) = \\
G(A)(1 - G(A))/(1 + a). \)

A key feature of the Dirichlet process is the conjugacy property. Specifically,
if \( x = (x_1, \ldots, x_n) \) is a random sample from \( P \sim DP(a,G) \), then the posterior
distribution of \( P \) is \( P | x = P_x \sim DP(a + n, G_x) \), where

\[
G_x = a(a + n)^{-1}G + n(a + n)^{-1}F_n
\] (5)
with $F_i(\cdot) = n^{-1} \sum_{i=1}^{n} \delta_{x_i}(\cdot)$ and $\delta_{x_i}$ is the Dirac measure at $x_i$ (i.e. $\delta_{x_i}(B) = 1$ if $x_i \in B$ and 0 otherwise). Notice that $G_x$ is a convex combination of the prior base distribution $G$ and the empirical distribution $F_n$. Clearly, $G_x \to G$ as $a \to \infty$ while $G_x \to F_n$ as $a \to 0$. On the other hand, by the Glivenko-Cantelli theorem, when $a/n \to 0$ (i.e., $a$ is small comparable to $n$), $G_x$ converges almost surely to the true distribution function. We refer the reader to Ishwaran and James (2001), Al-Labadi and Zarepour (2014) and Al-Labadi and Abdelrazeq (2017) for other asymptotic properties of the Dirichlet process.

Ferguson (1973) proposed an alternative definition of the Dirichlet process via a series representation. Specifically, if $P \sim DP(a, G)$, then

$$P(\cdot) = \sum_{i=1}^{\infty} L^{-1}(\Gamma_i) \delta_{Y_i}(\cdot) / \sum_{i=1}^{\infty} L^{-1}(\Gamma_i),$$

(6)

where $\Gamma_i = E_1 + \cdots + E_i$, $E_i \overset{i.i.d.}{\sim} \text{exponential}(1)$, $Y_i \overset{i.i.d.}{\sim} G$ independent of $\Gamma_i$, and $L^{-1}(y) = \inf\{x > 0 : L(x) \geq y\}$ with $L(x) = \int_{x}^{\infty} t^{-1}e^{-t}dt, x > 0$. It follows from (6) that the Dirichlet process is a discrete probability measure even for the cases with an absolutely continuous base measure $G$. Despite this fact, by imposing the weak topology, the support of the Dirichlet process could be quite large, namely, the set of all probability measures whose support is contained in the support of the base measure.

Since working with (6) is not straightforward (inverse of $L(x)$ has no closed form), Ishwaran & Zarepour (2002) developed the following easy-to-use representation to simulate the Dirichlet process. They showed that $P \sim DP(a, G)$ can be approximated by

$$P_N(\cdot) = \sum_{i=1}^{N} J_{i,N} \delta_{Y_i}(\cdot),$$

(7)

where $(J_{1,N}, \ldots, J_{N,N}) \sim \text{Dirichlet}(a/N, \ldots, a/N)$. More precisely, $(P_N)_{N \geq 1}$
converges in distribution to $P$, where $P_N$ and $P$ are random values in the space $M_1(\mathbb{R})$ of probability measures on $\mathbb{R}$ endowed with the topology of weak convergence. To generate $(J_{i,N})_{1 \leq i \leq N}$ make $J_{i,N} = W_{i,N} / \sum_{i=1}^{N} W_{i,N}$, where $(W_{i,N})_{1 \leq i \leq N}$ is a sequence of i.i.d. gamma($a/N, 1$) random variables independent of $(Y_i)_{1 \leq i \leq N}$. An interesting feature of dealing with (7) is the exchangeability of the weights $(J_{i,N})_{1 \leq i \leq N}$ and the simplicity of dealing with the Dirichlet distribution. This form of approximation plays a central role in the following sections.

For an extensive discussion about various simulation methods of the Dirichlet process, consult Zarepour and Al-Labadi (2012) and references therein. Throughout the paper, the same notation will be used for the measure and its distribution function. That is, $P(t) = P((−∞, t]), t \in \mathbb{R}$.

3 Bayesian Estimation of the Entropy

Let $P_N(\cdot) = \sum_{i=1}^{N} J_{i,N} \delta_{Y_i}(\cdot)$. Let $m$ be a positive integer smaller than $N/2$, $Y_{(i−m)} = Y_{(1)}$ if $i \leq m$, $Y_{(i+m)} = Y_{(N)}$ if $i \geq N − m$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}$ are the order statistics of $Y_1, Y_2, \ldots, Y_N$. The slope of the straight line that joins the two points $(P_N(Y_{(i−m)}), Y_{(i−m)})$ and $(P_N(Y_{(i+m)}), Y_{(i+m)})$ is

$$\frac{Y_{(i+m)} - Y_{(i−m)}}{P_N(Y_{(i+m)}) - P_N(Y_{(i−m)})} = \frac{Y_{(i+m)} - Y_{(i−m)}}{c_{i,a}},$$

where

$$c_{i,a} = \begin{cases} \sum_{k=2}^{i+m} J_{k,N} & \text{if } 1 \leq i \leq m \\ \sum_{k=i-m+1}^{i+m} J_{k,N} & \text{if } m + 1 \leq i \leq N − m \\ \sum_{k=i-m+1}^{N} J_{k,N} & \text{if } N − m + 1 \leq i \leq N \end{cases},$$

(8)
For instance, if \(1 \leq i \leq m\), then
\[
c_i,a = P_N(Y_{(i+m)}) - P_N(Y_{(1)}) = \sum_{k=1}^{i+m} J_{k,N} - J_{1,N} = \sum_{k=2}^{i+m} J_{k,N}.
\]

Note that, from the properties of the Dirichlet distribution, \(J_{i,N} \sim \text{beta}(a/N, a(1 - 1/N))\). Thus, \(E(J_{i,N}) = N^{-1}\). Hence,
\[
E[c_{i,a}] = \begin{cases} 
\frac{m+i-1}{N} & \text{if } 1 \leq i \leq m \\
\frac{2m}{N} & \text{if } m + 1 \leq i \leq N - m \\
\frac{N+m-i}{N} & \text{if } N - m + 1 \leq i \leq N
\end{cases}
= \frac{m}{N} c_i,
\]

where \(c_i\) is defined in (4) with \(n\) is replaced by \(N\). The next proposition underlines a direct connection between \(c_{i,a}\) and \(c_i\). Its proof is given in the Appendix.

**Proposition 1** Let \((J_1,N, J_2,N, \ldots, J_{N,N}) \sim \text{Dirichlet}(a/N, \ldots, a/N)\). As \(N \to \infty\),

1. \(J_{i,N} - 1/N \overset{P}{\to} 0\)
2. \(c_{i,a} - \frac{m}{N} c_i \overset{P}{\to} 0\), where \(c_{i,a}\) and \(c_i\) are defined in (8) and (4), respectively.

Proposition 1 is essential for constructing a Bayesian nonparametric estimator to entropy based on the Dirichlet process. The form of the prior of this estimator is presented in the next lemma. The proof is placed in the Appendix.

**Lemma 2** Let \(P_N = \sum_{i=1}^{N} J_{i,N} \delta_{Y_i}\) as defined in Section 2, where \(Y_1, Y_2, \ldots, Y_N \overset{i.i.d.}{\sim} G\). Let \(m\) be a positive integer smaller than \(N/2\), \(Y_{(i-m)} = Y_{(1)}\) if \(i \leq m\), \(Y_{(i+m)} = Y_{(N)}\) if \(i \geq N - m\) and \(Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}\) are the order statistics
of $Y_1, Y_2, \ldots, Y_N$. Let

$$H_{m,N,a} = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{Y_{(i+m)} - Y_{(i-m)}}{c_{i,a}} \right),$$ (10)

where $c_{i,a}$ is defined in (8). As $N \to \infty$, $m \to \infty$, $m/N \to 0$ and $a \to \infty$, we have

$$E[H_{m,N,a}] - E[H_{m,n}^{EPS}] \to 0,$$

where $H_{m,n}^{EPS}$ is defined in (3).

The next lemma shows that the estimator defined in (10) is consistent. A formal proof is given in the Appendix.

**Lemma 3** Let $H_{m,N,a}, N, m, a$ and $G$ be as defined in Lemma 2. Then as $N \to \infty$, $m \to \infty$, $m/N \to 0$ and $a \to \infty$, we have

$$H_{m,N,a} \xrightarrow{p} H(G) = -\int_{-\infty}^{\infty} g(x) \log g(x)dx,$$

where $G'(x) = g(x)$.

The following lemma gives the posterior version $H_{m,N,a|x}$ of the prior $H_{m,N,a}$ and addresses its consistency. The proof follows from (5), the Glivenko-Cantelli theorem and Lemma 3.

**Lemma 4** Let $x = (x_1, \ldots, x_n)$ be a sample from $F$. Let $H_{m,N,a|x}$ be the posterior version of $H_{m,N,a}$ as defined in (10) with $P_N$ replaced by $P_N|x$, where $P_N|x$ is an approximation of $P|x \sim DP(a + n, G_x)$. Then as $N \to \infty$, $m \to \infty$, $n \to \infty$, $m/N \to 0$ and $a/n \to 0$, we have

$$H_{m,N,a|x} \xrightarrow{p} H(F) = -\int_{-\infty}^{\infty} f(x) \log f(x)dx.$$
4 Choices of $m$, $a$ and $G$

Let $x = (x_1, \ldots, x_n)$ be a sample from a continuous distribution $F$. The aim is to approximate $H(F)$ via the approximation in Lemma 4. To proceed with this approximation, it is necessary to discuss the choices for $m$, $a$ and $G$. We start by the choice of $m$. It is common to use the following formula due to Grzegorzewski and Wieczorkowski (1999):

$$m = \lfloor \sqrt{N} + 0.5 \rfloor,$$

(11)

where $\lfloor y \rfloor$ is the largest integer less than or equal to $y$. For instance, for $N = 10, 20, 50$, the choices of $m$ using (11) are 3, 4, 7, respectively. Note that, the value of $m$ in (11) is the value that will be used for the prior. For the posterior, one should replace $N$ by the number of distinct atoms in $P_N|x$, an approximation of $F|x$. Notice that, from (5), if $a/n$ is close to zero, then the number of distinct atoms in $P_N|x$ will be roughly $n$.

As for the hyperparameters $a$ and $G$ of the Dirichlet process, their choices depend on the application of interest. For entropy estimation, any choice of $a$ such that $a/n$ is close to zero should work for any choice of $G$. This follows from (5) as when $a/n$ is close to 0, the sample will dominate the prior guess $G$. That is, the approach becomes invariance to any choice of $G$. For example, setting $a = 0.05$ and $n = 10$ in (5) gives

$$G_x = 0.005G + 0.995F_n.$$

This implies the chance to draw a sample from the collected data is 99.5% over a new sample from $G$. As a demonstrative purpose, we will set $G = N(0,1)$ and $a = 0.05$, although other choices are certainly possible. An example studying the sensitivity of the approach to the choice of $G$ is covered in Section 4, Table
Now, based on Lemma 4, the following computational algorithm is proposed for estimating (1).

**Algorithm A (Nonparametric Estimation of Entropy)**

(i) Let \( P \sim DP(a, G) \) and \( P_N \) be an approximation of \( P \). Set \( a = 0.05 \) and \( G = N(0, 1) \).

(ii) Generate a sample from \( P_N|x \), where \( P_N|x \) is an approximation of \( P|x \sim DP(a + n, G_x) \). See Section 2.

(iii) Compute \( H_{m,N,a}|x \) as in Lemma 4.

(iv) Repeat steps (i)-(iii) to obtain a sample of \( r \) values from \( H_{m,N,a}|x \). For large \( r \), the empirical distribution of these values is an approximation to the distribution of \( H_{m,N,a}|x \).

(v) The average of the \( r \) values generated in step (iv) will be the estimator of the entropy.

5 Examples

In this section, we study the behaviour of the proposed estimator in terms of efficiency and robustness. The proposed estimator is compared with its non-Bayesian counterpart estimators of Vasicek (1976) and Ebrahimi, Pflughoeft and Soofi (1994). Additionally, for a comprehensive comparison, we included the (weighted) Kozachenko–Leonenko (KL) entropy estimator (Kozachenko and Leonenko, 1987; Berrett, Samworth and Yuan, 2019). This estimator is based on the \( k \)-nearest neighbour distances of the sample. The comparison between Bayesian and non-Bayesian methods makes sense here as setting \( a = 0.05 \) produces an estimator that is independent from any choice of the prior guess \( G \). For
each sample size \((n = 10, 20, 50)\), 1000 samples were generated. We have considered four distributions: uniform on \((0, 1)\) (exact entropy is 0), exponential with mean 1 (exact entropy is 1), \(N(0, 1)\) (exact entropy is \(0.5 \log(2\pi e) \approx 1.419\)) and Weibull distribution with shape parameter equal to 2 and scale parameter equal to 0.5 (exact value of \(-0.098\)). The estimators and their root mean squared errors are computed and reported in Table 1-Table 5. Here each sample of the 1000 samples gives an estimate. The reported value of the estimator (Est) is the average of the 1000 estimates. In Table 2 we have reported the median instead of the mean for the uniform(0, 1) case. On the other hand, the root mean squared error (RMSE) is computed as follows: 

\[
\sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\hat{H}_i(F) - \text{true value})^2}
\]

where \(\hat{H}_i(F)\) is the estimated value based on the ith sample. The computing program codes were implemented in the programming language R and are available from the authors. In Algorithm A, we set \(r = 1000\) and \(N = 200\). For the KL entropy estimator, we used the package IndepTest (Berrett, Grose and Samworth, 2018) with \(k = m\) as in formula (11).

Table 1: uniform(0, 1). Exact value is 0. Est and RMSE stand for the estimated value and the root mean squared error.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(m)</th>
<th>(H_{m,N,a})</th>
<th>(H_{m,n,a}^V)</th>
<th>(H_{m,n,a}^{EPS})</th>
<th>KL Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
<td>-0.037(0.158)</td>
<td>-0.423(0.455)</td>
<td>-0.167(0.235)</td>
<td>0.044(0.286)</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>0.004(0.084)</td>
<td>-0.255(0.270)</td>
<td>-0.097(0.130)</td>
<td>0.037(0.182)</td>
</tr>
<tr>
<td>50</td>
<td>7</td>
<td>0.013(0.032)</td>
<td>-0.153(0.158)</td>
<td>-0.053(0.063)</td>
<td>0.007(0.114)</td>
</tr>
</tbody>
</table>

Figure 1 represents boxplots of 1000 estimated values of entropy. Each value is computed from a different sample generated from the uniform(0, 1) distribution. Here we have used the same samples used in Table 1. It follows clearly from the Figure 1 the existence of outliers in the estimated values. Thus, one
may consider the median instead of the mean as an estimate of entropy. Table 2 presents estimates of the entropy of the uniform(0,1) distribution based on the median.

![Boxplots](image)

(a) $n = 10$

(b) $n = 20$

(c) $n = 50$

Figure 1: Boxplots of 1000 estimated values of entropy. Each value is computed from a sample of size $n$, where $n = 10, 20, 50$. 
Table 2: uniform(0, 1). Exact value is 0. Est and RMSE stand for the estimated value and the root mean squared error.

| n  | m  | \(H_{m,N,a}|x\) Est(RMSE) | \(H^V_{m,n,a}\) | \(H^{EPS}_{m,n,a}\) | KL Entropy  |
|----|----|----------------------------|-----------------|-----------------|-------------|
| 10 | 3  | -0.005                     | -0.388          | -0.131          | 0.077       |
| 20 | 4  | 0.019                       | -0.241          | -0.082          | 0.053       |
| 50 | 7  | 0.019                       | -0.148          | -0.048          | 0.011       |

Table 3: exponential with mean 1. Exact value is 1. Est and RMSE stand for the estimated value and the root mean squared error.

| n  | m  | \(H_{m,N,a}|x\) Est(RMSE) | \(H^V_{m,n,a}\) Est(RMSE) | \(H^{EPS}_{m,n,a}\) Est(RMSE) | KL Entropy Est(RMSE) |
|----|----|----------------------------|----------------------------|-------------------------------|----------------------|
| 10 | 3  | 0.996(0.313)               | 0.562(0.567)              | 0.818(0.402)                  | 0.929(0.405)         |
| 20 | 4  | 1.022(0.228)               | 0.752(0.344)              | 0.911(0.255)                  | 0.963(0.276)         |
| 50 | 7  | 1.030(0.145)               | 0.880(0.187)              | 0.980(0.145)                  | 0.979(0.155)         |

Table 4: \(N(0, 1)\). Exact value is 1.419. Est and RMSE stand for the estimated value and the root mean squared error.

| n  | m  | \(H_{m,N,a}|x\) Est(RMSE) | \(H^V_{m,n,a}\) Est(RMSE) | \(H^{EPS}_{m,n,a}\) Est(RMSE) | KL Entropy Est(RMSE) |
|----|----|----------------------------|----------------------------|-------------------------------|----------------------|
| 10 | 3  | 1.222(0.308)               | 0.866(0.611)              | 1.122(0.395)                  | 1.285(0.340)         |
| 20 | 4  | 1.302(0.207)               | 1.087(0.375)              | 1.246(0.247)                  | 1.342(0.232)         |
| 50 | 7  | 1.389(0.110)               | 1.259(0.192)              | 1.359(0.122)                  | 1.398(0.145)         |
Table 5: Weibull with shape parameter 2 and scale parameter 0.5. Exact value is $-0.098$. Est and RMSE stand for the estimated value and the root mean squared error, respectively.

| $n$ | $m$ | $H_{m,N,a}|x$ (Est(RMSE)) | $H_{m,n,a}^V$ (Est(RMSE)) | $H_{m,n,a}^{EPS}$ (Est(RMSE)) | KL Entropy (Est(RMSE)) |
|-----|-----|--------------------------|---------------------------|-------------------------------|------------------------|
| 10  | 3   | -0.020(0.212)            | -0.634(0.593)             | -0.377(0.378)                | -0.198(0.339)          |
| 20  | 4   | -0.108(0.152)            | -0.431(0.373)             | -0.273(0.242)                | -0.173(0.245)          |
| 50  | 7   | -0.102(0.100)            | -0.260(0.192)             | -0.160(0.118)                | -0.114(0.148)          |

It follows clearly from Table 1 - Table 5 that, the new approximation of entropy has the lowest root mean squared error for most cases covered in this section.

It is also interesting to assess the sensitivity of the proposed estimation method to the choice of $a$ and $G$. To this end, we fixed $a$ at 0.05 and 5 with several values of $G$. The next data set generated from the exponential distribution with mean 20 is used in the study.


The results of the estimated entropy for the previous data set are reported in Table 6, where $N(\mu,\sigma^2)$ denotes the normal distribution with mean $\mu$ and standard deviation $\sigma$ and $t_1$ denotes the $t$ distribution with 1 degree of freedom.

It follows from Table 6 that when $a = 0.05$, then $G$ has no practical influence on the estimated entropy. On the other side, for large values of $a$ (such as $a = 5$), the estimated value depends on the choice of $G$. Therefore, it is recommended to set $a = 0.05$ and $G = N(0,1)$ in the proposed approach.
Table 6: Study of the effect of the proposed estimator using different values of $a$ and $G$.

<table>
<thead>
<tr>
<th>$G$</th>
<th>Estimate: $a = 0.05$</th>
<th>Estimate: $a = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(0, 1)$</td>
<td>3.491</td>
<td>3.126</td>
</tr>
<tr>
<td>$N(3, 9)$</td>
<td>3.492</td>
<td>3.512</td>
</tr>
<tr>
<td>$t_1$</td>
<td>3.501</td>
<td>3.910</td>
</tr>
<tr>
<td>exponential(1)</td>
<td>3.476</td>
<td>2.667</td>
</tr>
<tr>
<td>uniform(0, 1)</td>
<td>3.474</td>
<td>2.354</td>
</tr>
</tbody>
</table>

6 Conclusion

In this paper, an efficient, yet simple, Bayesian nonparametric estimator of entropy is proposed. The proposed estimator is considered an analogous Bayesian estimator to the estimator of Ebrahimi, Pflughoeft and Soofi (1994). Through several examples, it has been shown that the approach performs extremely well where a smaller root mean squared error is obtained. The foremost motive of having this estimator is to use it in applications such as model checking as discussed, for instance, in Al-Labadi and Evans (2018). We have left this critical avenue of research to future work.

Acknowledgment

The authors are grateful to the Editor, the Associate Editor and the reviewer for their comments and suggestions that have greatly improved the paper.
References


A Proofs

A.1 Proof of Proposition 1

1. For each $1 \leq i \leq N$, from the properties of the Dirichlet distribution, $J_{i,N} \sim \text{beta}(a/N, a(1 - 1/N))$. It follows that $E[J_{i,N}] = 1/N$ and

$$V[J_{i,N}] = \frac{1/N(1 - 1/N)}{a + 1},$$

where $V$ stands for the variance. Since, as $N \to \infty$, $V(J_{i,N}) \to 0$, we conclude the result.

2. By (9), $E[c_{i,a}] = \frac{m}{N}c_i$. Thus, to prove the proposition, by Chebyshev’s inequality, it is sufficient to show that $V(c_{i,a}) \to 0$. We consider three cases.

Case I (for $1 \leq i \leq m$): From the aggregation property of the Dirichlet distribution,

$$\sum_{k=2}^{i+m} J_{k,N} \sim \text{beta} \left( \sum_{k=2}^{i+m} \frac{a}{N}, a - \sum_{k=2}^{i+m} \frac{a}{N} \right). \quad (12)$$

Hence, as $N \to \infty$,

$$V \left( \sum_{k=2}^{i+m} J_{k,N} \right) = \frac{\sum_{k=2}^{i+m} \frac{a}{N} \left( a - \sum_{k=2}^{i+m} \frac{a}{N} \right)}{a^2(1 + a)}$$

$$= \frac{(i + m - 1)(N - i - m + 1)}{N^2(a + 1)} \to 0.$$

Case II (for $m + 1 \leq i \leq N - m$): similar to Case I,

$$\sum_{k=i-m+1}^{i+m} J_{k,N} \sim \text{beta} \left( \sum_{k=i-m+1}^{i+m} \frac{a}{N}, a - \sum_{k=i-m+1}^{i+m} \frac{a}{N} \right). \quad (13)$$
Hence, as $N \to \infty$,

$$V \left( \sum_{k=i-m+1}^{i+m} J_{k,N} \right) = \frac{\sum_{k=i-m+1}^{i+m} \frac{a}{N} \left( a - \sum_{k=i-m+1}^{i+m} \frac{a}{N} \right)}{a^2(a+1)}$$

$$= \frac{2m(N-2m)}{N^2(a+1)} \to 0.$$

Case III (for $N-m+1 \leq i \leq N$): As in the previous cases,

$$\sum_{k=i-m+1}^{N} J_{k,N} \sim \text{beta} \left( \sum_{k=i-m+1}^{N} \frac{a}{N}, a - \sum_{k=i-m+1}^{N} \frac{a}{N} \right). \quad (14)$$

Therefore, as $N \to \infty$,

$$V \left( \sum_{k=i-m+1}^{N} J_{k,N} \right) = \frac{\sum_{k=i-m+1}^{N} \frac{a}{N} \left( a - \sum_{k=i-m+1}^{N} \frac{a}{N} \right)}{a^2(1+a)}$$

$$= \frac{(N-i+m)(i-m)}{N^2(a+1)} \to 0.$$

This completes the proof of the proposition. ■

A.2 Proof of Lemma 2

Recall that,

$$H_{m,N,a} = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{Y_{(i+m)} - Y_{(i-m)}}{c_{i,a}} \right)$$

and

$$H_{m,N}^{EPS} = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{Y_{(i+m)} - Y_{(i-m)}}{mc_{i}/N} \right),$$
where \( c_{i,a} \) and \( c_i \) are defined, respectively, on (8) and (4). Thus,

\[
E[H_{m,N,a}] - E[H_{EPS}^{m,N}] = \frac{1}{N} \sum_{i=1}^{N} E\left[ \log \left( \frac{mc_i/N}{c_{i,a}} \right) \right] = \frac{1}{N} \sum_{i=1}^{N} \log (c_i m/N) - \frac{1}{N} \sum_{i=1}^{N} E[\log c_{i,a}]. \tag{15}
\]

We want to show that, as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \), (15) \( \to 0 \).

We consider three cases.

**Case I** (for \( 1 \leq i \leq m \)): notice that, with \( \alpha_k = aN^{-1} \) and \( \alpha_0 = \sum_{k=1}^{N} \alpha_k = a \),

\[
E[J_{i,N} \log c_{i,a}] = E\left[ J_{i,N} \log \left( \sum_{k=2}^{i+m} Z_{k,N} \right) \right] = \\
\int \ldots \int z_i \log \left( \sum_{k=2}^{i+m} z_k \right) \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{N} \Gamma(\alpha_k)} z_1^{\alpha_0-1} \ldots z_i^{\alpha_i-1} \ldots z_N^{\alpha_N-1} dz_1 \ldots dz_i \ldots dz_N, \\
= \int \ldots \int \log \left( \sum_{k=2}^{i+m} z_k \right) \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{N} \Gamma(\alpha_k)} z_1^{\alpha_i-1} \ldots z_i^{(\alpha_0+1)-1} \ldots z_N^{\alpha_N-1} dz_1 \ldots dz_i \ldots dz_N \\
= \frac{\alpha_i}{\alpha_0} E\left[ \log \left( \sum_{k=2}^{i+m} Z_{k,N} \right) \right], \tag{16}
\]

where \( \sum_{k=2}^{i+m} Z_{k,N} \sim \text{beta} \left( \sum_{k=2}^{i+m} \alpha_k + 1, (a + 1) - \left( \sum_{k=2}^{i+m} \alpha_k + 1 \right) \right) \). For \( \alpha_k = aN^{-1} \), \( \sum_{k=2}^{i+m} Z_{k,N} \sim \text{beta} \left( a(m + i - 1)N^{-1} + 1, a - a(m + i - 1)N^{-1} \right) \). From the properties of the beta distribution, we have

\[
(16) = \frac{\alpha_i}{\alpha_0} \left( \psi \left( \sum_{k=1}^{m} \alpha_k + 1 \right) - \psi (\alpha_0 + 1) \right) \\
= \frac{1}{N} \left( \psi \left( \frac{a(m + i - 1)}{N} + 1 \right) - \psi (a + 1) \right), \tag{17}
\]

where \( \psi(x) = \Gamma'(x)/\Gamma(x) \) is the digamma function. Therefore, by (17) and for
1 \leq i \leq m, we obtain

\begin{equation}
(15) = \frac{1}{N} \sum_{i=1}^{m} \log \left( \frac{m + i - 1}{N} \right) - \frac{1}{N} \sum_{i=1}^{m} \left( \psi \left( \frac{a(m + i - 1)}{N} + 1 \right) - \psi(a + 1) \right).
\end{equation}

Using that facts that \( \psi(x + 1) = \log(x) + O(x^{-1}) \) and \( \sum_{i=0}^{L-1} \frac{1}{x+i} = \psi(x + L) - \psi(x) = \log \left( \frac{x+L}{2} \right) + O(x^{-1}) \) (Abramowitz and Stegun, 1972), we have

\begin{align*}
(15) &\approx - \frac{1}{N} \sum_{i=1}^{m} O \left( \frac{N}{a(m + i - 1)} \right) + \frac{1}{N} \sum_{i=1}^{m} O \left( \frac{1}{a} \right) \\
&= \frac{1}{a} O \left( \sum_{i=1}^{m} \frac{1}{a(m + i - 1)} \right) + O \left( \frac{m}{Na} \right) \\
&= \frac{1}{a} O \left( \psi(2m) - \psi(m - 1) \right) + O \left( \frac{m}{Na} \right) \\
&= \frac{1}{a} O \left( \log \left( \frac{2m}{m-1} \right) + \frac{1}{2m} \right) + O \left( \frac{m}{Na} \right) \to 0,
\end{align*}

as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \).

Case II (for \( m + 1 \leq i \leq N - m \)): similar to Case I,

\begin{equation}
E \left[ J_{i,N} \log c_{i,a} \right] = \frac{\alpha_i}{\alpha_0} E \left[ \log \left( \sum_{k=i-m+1}^{i+m} Z_{k,N} \right) \right], \quad (18)
\end{equation}

where \( \sum_{k=i-m+1}^{i+m} Z_{k,N} \sim \text{beta}(2amN^{-1} + 1, a - 2amN^{-1}) \). Thus, from the properties of the beta distribution, we have

\begin{equation}
(18) = \frac{1}{N} \left( \psi \left( \frac{2am}{N} + 1 \right) - \psi(a + 1) \right), \quad (19)
\end{equation}

Therefore, by (19), we have
\[ (15) = \sum_{i=m+1}^{N-m} \log \left( \frac{2m}{N} \right) - \frac{1}{N} \sum_{i=m+1}^{N-m} \left( \psi \left( \frac{2am}{N} + 1 \right) + \psi(a + 1) \right) \]
\[ = O \left( \frac{N}{2am} \right) + O \left( \frac{1}{a} \right) \to 0, \]
as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \).

**Case III** (for \( N - m + 1 \leq i \leq N \)): similar to the previous cases,
\[ E[J_{i,N} \log c_{i,a}] = \frac{\alpha_i}{\alpha_0} E \left[ \log \left( \sum_{k=i-m+1}^{N} Z_{k,N} \right) \right], \quad (20) \]
where \( \sum_{k=i-m+1}^{N} Z_{k,N} \sim \text{beta} \left( a(N + m - i)N^{-1} + 1, a - a(N + m - i)N^{-1} \right) \).

Thus, from the properties of the beta distribution, we have
\[ (20) = \frac{1}{N} \left( \psi \left( \frac{a(N + m - i)}{N} + 1 \right) - \psi(a + 1) \right). \quad (21) \]

Therefore, by (21), we have
\[ (15) = \frac{1}{N} \sum_{i=N+m+1}^{N} \log \left( \frac{N + m - i}{N} \right) \]
\[ - \frac{1}{N} \sum_{i=N-m+1}^{N} \psi \left( \frac{a(N + m - i)}{N} + 1 \right) + \psi(a + 1) \]
\[ = O \left( \sum_{i=N-m+1}^{N} \frac{1}{a(N + m - i)} \right) + O \left( \frac{m}{aN} \right) \]
\[ = O \left( \frac{\psi(1 - 2m)}{a} - \frac{\psi(1 - m)}{a} \right) + O \left( \frac{m}{aN} \right) \to 0, \]
as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \). Thus, in all cases, as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \), (15) \( \to 0 \). This completes the proof of Lemma 2. ■
A.3 Proof of Lemma 3

Note that,

\[ H_{m,N,a} = (H_{m,N,a} - H_{m,N}^{EPS}) + H_{m,N}^{EPS}, \]

where \( H_{m,N}^{EPS} \) is the approximation given in (3). It follows that,

\[ H_{m,N,a} - H_{m,N}^{EPS} = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{mc_i/N}{c_i,a} \right). \]

Since \((J_{i,N})_{1 \leq i \leq N}\) is a sequence of pairwise negative associated identically distributed random variables with finite expectations, by Theorem 4.2.8 of Atkinson (2017), the weak law of large numbers holds for the sequence \(\left( \frac{mc_i/N}{c_i,a} \right)_{1 \leq i \leq N}\). Thus we have

\[ H_{m,N,a} - H_{m,N}^{EPS} - E \left( \log \left( \frac{mc_i/N}{c_i,a} \right) \right) \rightarrow 0. \]

To show that \(E \left( \log \left( \frac{mc_i/N}{c_i,a} \right) \right) \rightarrow 0\), we consider three cases.

**Case I (for 1 \leq i \leq m):** From (12) and the well-known property of the beta distribution, we have

\[ E \left( \log \left( \frac{mc_i/N}{c_i,a} \right) \right) = \log \left( \frac{mc_i}{N} \right) - E \left( \log (c_i,a) \right) \]
\[ = \log \left( \frac{i + m - 1}{N} \right) - \psi \left( \frac{a(i + m - 1)}{N} \right) + \psi(a) \]
\[ = \log \left( \frac{i + m - 1}{N} \right) - \log \left( \frac{a(i + m - 1)}{N} \right) - O \left( \frac{N}{a(i + m - 1)} \right) \]
\[ + \log(a) + O(\frac{1}{a}) \]
\[ = -O \left( \frac{N}{a(i + m - 1)} \right) + O(\frac{1}{a}) \rightarrow 0, \]

as \(N \rightarrow \infty\), \(m \rightarrow \infty\), \(m/N \rightarrow 0\) and \(a \rightarrow \infty\).
Case II (for \( m + 1 \leq i \leq N - m \)): From (13) and the well-known property of the beta distribution, we have

\[
E \left( \log \left( \frac{mc_i/N}{c_{i,a}} \right) \right) = \log \left( \frac{mc_i}{N} \right) - E \left( \log (c_{i,a}) \right)
\]

\[
= \log \left( \frac{2m}{N} \right) - \psi \left( \frac{2am}{N} \right) + \psi(a)
\]

\[
= \log \left( \frac{2m}{N} \right) - \log \left( \frac{2am}{N} \right) - O \left( \frac{N}{2am} \right)
\]

\[
+ \log(a) + O \left( \frac{1}{a} \right)
\]

\[
= -O \left( \frac{N}{2am} \right) + O \left( \frac{1}{a} \right) \to 0,
\]

as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \).

Case III (for \( N - m + 1 \leq i \leq N \)): As in the previous cases, from (14), we have

\[
E \left( \log \left( \frac{mc_i/N}{c_{i,a}} \right) \right) = \log \left( \frac{mc_i}{N} \right) - E \left( \log (c_{i,a}) \right)
\]

\[
= \log \left( \frac{n + m - i}{N} \right) - \psi \left( \frac{a(n + m - i)}{N} \right) + \psi(a)
\]

\[
= \log \left( \frac{n + m - i}{N} \right) - \log \left( \frac{a(n + m - i)}{N} \right) - O \left( \frac{N}{a(n + m - i)} \right)
\]

\[
+ \log(a) + O \left( \frac{1}{a} \right)
\]

\[
= -O \left( \frac{N}{a(n + m - i)} \right) + O \left( \frac{1}{a} \right) \to 0.
\]

as \( N \to \infty, m \to \infty, m/N \to 0 \) and \( a \to \infty \). Thus, in all cases, \( H_{m,N,a} - H_{m,N}^{EPS} \to 0 \). Also, by Ebrahimi, Pflughoeft and Soofi (1994), as \( N \to \infty, m \to \infty \) and \( m/N \to 0 \) we have \( H_{m,N}^{EPS} \to H(G) \). Now, applying Slutsky’s theorem (Ferguson, 1996) completes the proof. \( \blacksquare \)